



National Snow and Ice Data Center
Supporting Cryospheric Research Since 1976



Unique Identifiers Assessment: Results

R. Duerr

Outline

- Background
 - Identifier schemes
 - Assessment criteria
 - Levels of data
 - Use cases
- Assessment Results

Identifier schemes assessed

- Archival Resource Key (ARK)
- Digital Object Identifiers (DOI)
- Extensible Resource Identifier (XRI)
- HANDLE
- Life Science ID (LSID)
- Object Identifiers (OID)
- Persistent Uniform Resource Locators (PURL)
- URI/URN/URL

Assessment Criteria

- Technical value (Standard? Security? Scalability? Interoperability? Internet compatibility? 3rd party maintenance? Naming authority and stability? Expected longevity?)
- User value (Usable in citations? Any additional trust value? Opaque or transparent?)
- Archive value (Costs, Ease of migration, Extensible to non-web based objects, physical objects?)
- Existing usage within data centers

Data Levels

- Only two levels of data identifier are addressed:
 - A data set as a whole
 - Individual files or granules within a data set

Use Case #1

- To uniquely & unambiguously identify a particular piece of data no matter which copy a user has
- Ideal attributes
 - Location independent (i.e., copies everywhere have this same ID)
 - Generate at time of object creation
 - Placeable inside the object or it's metadata
- Practical attributes

Use Case #2

- To locate the authoritative copy of the data no matter where they are currently held
- Ideal attributes
 - Location invariant (i.e., no matter where the master copy moves, this ID remains the same and can always be used to find it)
 - Globally unique
 - Resolves to a Use Case #1 ID
- Practical attributes
 - External name authority necessary
 - Generate only on decision to make data permanently available

Use Case #3

- To identify the data cited in a particular publication
- Ideal attributes
 - Basically those of Use Case #2 with a couple of caveats
 - Acceptance by publishers and authors
 - Facilitate identification at the data set or data set subset level

Use Case #4

- To be able to tell that two files contain the same data even if the formats are different. In other words, to determine if two files are “scientifically identical” to use Curt Tilmes' terminology.
- Ideal Attributes
 - Same as Use Case #1 plus
 - Possible to verify that the contents are unchanged after

Assessment Results

ARK - Summary

- Form [http://NMAH/]ark:/NAAN/Name[Qualifier]
- Example: <http://ark.cdlib.org/ark:/13030/tf5p30086k>
- Fully qualified ARK's are URL's with added trust value:
 - ? Qualifier provides metadata
 - ?? Qualifier provides commitment statement
- Roughly two dozen name authorities exist including Google and the Internet Archive

ARK - Use Cases Supported

- Use Case 1 – supports all ideal attributes and most practical attributes for both data sets and data files. Does require a name authority.
- Use Case 2 – fully supported at both data levels
- Use Case 3 – not explicitly supported by publishers

DOI - Summary

- Form doi:[prefix]/[suffix]
- E.g., doi:10.3334/ORNLDAAAC/840
- DOI's are a type of Handle and URI
- ANSI/NISO standard Z39.84-2005
- DOI's can incorporate other pre-existing identifiers

DOI - Use Cases Supported

- Use Case 1 – not location independent
- Use Case 2 – fully supported for data sets but per DOI costs do not scale for data files
- Use Case 3 – fully supported
- Use Case 4 – not supported

XRI - Summary

- E.g., @nsidc.org+dataset*newName!(doi:10.12345/OriginalName)
- Has a relatively complex scheme that can represent quite a few concepts including the concept of a defining authority
 - = human authority
 - @ organizational authority
 - + dictionary concept for authority
 - \$ standards organization is the authority
- Transport protocol independent (e.g., http, FTP, xmpp)

XRI - Use Cases Supported

- Use Case 1 – not location independent
- Use Case 2 – fully supported and might survive past the end of http
- Use Case 3 – not familiar with either authors or publishers
- Use Case 4 – not supported

Handle - Summary

- Form [naming authority]/[local name]
- E.g., 10.1045 /january99–bearman
- CRNI manages the Global Handle Registry of name authorities
- Handles can incorporate other pre-existing identifiers

Handles - Use Cases Supported

- Use Case 1 – not location independent
- Use Case 2 – fully supported
- Use Case 3 – does not have the buy in with publishers that DOI's do and users are less familiar with them
- Use Case 4 – not supported

LSID - Summary

- Format: URN:LSID:<Authority>:<Namespace>:<ObjectID>[:<Version>]
- Example: urn:lsid:ncbi.nlm.nih.gov:GenBank:T48601:2
- Uniquely identify entities of interest to the life sciences
- Controversy over whether LSIDs violates the principle of reusing existing URI schemes
- This controversy apparently seems to have stopped development

OID - Summary

- Format: #.#.#...
- Example: 1.3.6.1 is the OID for the Internet
- OID's can be obtained from IANA and ANSI (among others)
- Authority to create new numbers past some point can be delegated
- It is not immediately obvious what to do with an OID if data transfers ownership

OID - Use Cases Supported

- Use Case 1 – not location independent
- Use Case 2 – not location invariant either; would require development of a resolution service
- Use Case 3 – does not have the buy in with publishers that DOI's do and users are not familiar with them

PURL - Summary

- Format: [protocol][resolver address][domain][name]
- Example: <http://purl.oclc.org/NET/EMILLER>
- Been around for more than a decade
- IETF and W3C support
- Recently upgraded to support a PURL federation of resolvers
- Added support for semantic concepts such as people, organizations, concepts, and data

PURL - Use Cases Supported

- Use Case 1 – not location independent
- Use Case 2 – fully supported
- Use Case 3 – does not have the buy in with publishers that DOI's do and users are not familiar with them
- Use Case 4 – not supported

URL/URN/URI - Summary

- IETF maintains the URI specification
- URL's require domain name purchase and are well known for their impermanence
- Redirection required if the object moves
- URN's do not necessarily imply the existence of a resource and a resolution service would need to be established

URI/URL/URN - Use Cases Supported

- Use Case 1 – URL's not location independent; URN's require registration
- Use Case 2 – URL's fully support
- Use Case 3 – does not have the buy in with publishers that DOI's do
- Use Case 4 – not supported

UUID - Summary

- Example: 0a9ecf4f-ab79-4b6b-b52a-1c9d4e1bb12f
- ISO/IEC 1578:1996 and IETF RFC 4122
- Intended to allow unique identification of objects by distributed systems without coordination
- Hash generation techniques can have additional trust value as they serve to indicate that the content has not been altered

UUID - Use Cases Supported

- Use Case 1 – Fully supported
- Use Case 2 – Not supported
- Use Case 3 – Not supported
- Use Case 4 – not supported

UUID - Use Cases Supported

Use Case	ARK	DOI	XRI	Hdl	LSID	OID	PURL	URL/ URI/ URN	UUID
#1	Yellow	Red							Green
#2	Green	Yellow	Green		Red			Green	Red
#3	Yellow	Green	Yellow		Red			Yellow	
#4	Red								

Best Practices

- Recognize that different identifier schemes are meant to solve different problems
- Recognize that a minimum of two identifiers will be needed for any data set or data file
- Plan for scheme obsolescence and replacement